Corporizando algumas questões*

*Diana Santos***

Resumo: Este artigo ambiciona: (i) apelar à consideração da língua portuguesa como entidade global, descrevendo algumas vantagens de ver a situação por este prisma; (ii) apresentar uma definição adequada de *corpo linguístico*, assim como uma tipologia de estudos com corpos; (iii) desenvolver questões metodológicas, esclarecendo algumas noções associadas; e (iv) apresentar alguns temas que me parecem de interesse para o futuro da área.
Palavras-chave: metodologia, língua portuguesa, linguística com corpos

Abstract: This paper (i) suggests the consideration of Portuguese *corpus* linguistics as the perfect arena to conceive Portuguese as international language; (ii) provides a definition of *corpus* and a typology of *corpus* linguistics; (iii) illustrates and discusses some methodological issues; and (iv) lists some interesting themes for future work.
Keywords: methodology, Portuguese, *corpus* linguistics

---

* Neste artigo foi mantida a grafia do português europeu em respeitro à origem da autora.

** Linguateca, SINTEF ICT, Oslo.

Preâmbulo

Muito agradeço a possibilidade que me dão de escrever um texto sobre a metodologia da linguística com corpos, algo que há muito me vem seduzindo, em vez da mera descrição de projectos ou recursos que tenha criado ou ajudado a criar. O meu objectivo é desencadear também nos leitores uma reflexão sobre a teoria e a prática do uso de corpos como ferramenta.

Aproveito também o ensejo para apelar à população de pesquisadores e linguistas brasileiros no sentido de uma visão do português como língua global e não apenas do Brasil. Observando o comportamento dos utilizadores do COMPARA, verifiquei que grande número de pedidos oriundos do Brasil selecciona apenas o português brasileiro, ao contrário dos outros utilizadores. Embora tal possa ser perfeitamente justificado no contexto particular de alguns desses pedidos, o seu número faz-me temer que represente uma atitude de desinteresse por outras variantes ou variedades da língua. O facto de ter consagrado os últimos dez anos da minha vida profissional à promoção do português como língua internacional, sem privilegiar uma variante específica (Santos, 1999, 2002) dá-me alguma autoridade moral para sugerir: vejamos a língua como algo que nos une e enriquece, e não como algo que nos separa, e isto sobretudo quando trabalhamos no estudo e processamento da própria língua.

Que o parágrafo acima não seja interpretado, jamais, como queixa dos meus colegas brasileiros: declaro que tenho sempre sido recebida de braços abertos no Brasil e, como se pode apreciar pela minha lista de publicações em português, exceptuando as que eu própria organizei ou editei, a esmagadora maioria foi publicada no Brasil, como aliás o presente artigo. Lamento que por razões conjunturais não me tenha sido possível estar em São Paulo em Setembro de 2007 no VI Encontro de LC e dizer o que aqui

escrevo de viva voz. Que fique bem clara a minha gratidão de poder publicar na minha língua, que aliás considero, como Matos (1992) argumenta, um direito linguístico inalienável e um dever como cientista.

Mas penso que todos concordarão em que é benéfico definir uma terminologia comum em português (afinal, a parte da língua em que temos algum controlo é a dos termos dos especialistas), e que olhar para texto parecido (neste caso, noutra variante) nos permite aguçar a sensibilidade linguística e a capacidade de observação, além de enriquecer o nosso potencial e a nossa criatividade como falantes. Porque, invocando por exemplo a actividade da tradução, quantas vezes o próprio tradutor não precisa de criar novos termos ou arranjar soluções criativas? Porque não "dar uma olhada" às eventualmente encontradas pelos seus colegas de além-mar? E, voltando à questão da terminologia, porque não enriquecer o português em conjunto em vez de o fragmentar? Vejamos o próprio exemplo da linguística com corpos: este último objecto tem sido variadamente chamado *corpora* (plural *corpora*), *córpus* (plural *córpora* ou *córpus*), mas parece não ter sido sequer equacionado o uso duma palavra genuinamente portuguesa e semelhante, *corpo*, empregue aliás de forma análoga em linguagem legal: *corpo de delito*. Na acepção mais lata de corpo como colecção de textos, é usada naturalmente a palavra *acervo* no Brasil, mas aparentemente não no sentido técnico associado a corpos electrónicos, mais influenciado pelo inglês. Infelizmente o uso não consagrou a possível expansão desse termo, provavelmente por não ter semelhanças suficientes com as designações inglesas/latinas. Proponho assim usar *corpo* e *corpos*, na esperança de que esta portuguesificação (e não aportuguesamento) seja aceite. Não por fazer questão em relação ao termo, mas porque me parece que a terminologia deva ser pensada e discutida pela comu-

nidade de especialistas numa área (e em todas as variantes) e não servilmente adoptada de outra língua.[1]

Material: o que é um corpo linguístico?

Após este preâmbulo, olhemos à nossa volta, em particular para a florescente área de linguística com corpos no Brasil, com mais de 75 artigos no VI Encontro e 40 grupos diferentes de pesquisa. Há dez anos, o interesse por corpos em português (e a qualidade e quantidade dos recursos, pelo menos os públicos) era mínimo, como o demonstra a panorâmica em Oksefjell e Santos (1998). A nível internacional, festejaram-se há pouco os 25 anos do ICAME (Facchinetti, 2007) com alguns artigos muito interessantes sobre a emergência do processamento de corpos por computador (embora com um pendor tipicamente anglo-saxónico/escandinavo) e, embora o ICAME seja, como o nome indica, dedicado ao estudo do inglês moderno (*International Computer Archive of Modern English*), muitos dos comentários e análises feitas no tal volume são, de facto, relevantes para qualquer língua.

Nesse contexto, pareceu-me pertinente salientar algumas das afirmações feitas na colectânea de artigos já mencionada e que me parecem estimulantes como objecto de reflexão. Uma das observações mais interessantes, feita por Svartvik (2007), foi a da própria mudança do estatuto dos corpos linguísticos, desde a altura em que um corpo era um objecto de valor que exigia e recomendava estudo

[1] O leitor atento terá certamente reparado que o título deste artigo ilustra o bónus que o uso de uma palavra genuinamente portuguesa implica, permitindo – na perspectiva de enriquecer a língua – vários novos sentidos de *corporizar*, *incorporar* ou mesmo *encorpar*, assim como a aplicação de sufixos produtivos como em *corpinho* ou *corpão*, *corpanzil*.

exaustivo (ou quase) – ou seja, o trabalho de compilar um corpo era tal que, depois, praticamente todos os fenómenos possíveis eram esmiuçados – até à presente proliferação e facilidade de obter todos os tipos de géneros de texto e de autores provocada pela Internet e patente na corrente da "Web como corpo" (Kilgarriff & Greffenstette, 2003).[2]

Em Santos (1998), defendi que um corpo electrónico, denominação vaga na altura, era de facto a conjunção de três coisas relacionadas: (i) um conjunto de textos, (ii) um conjunto de informação a marcar/classificar esses textos, e (iii) uma interface que permitisse consultar os dois primeiros. Por outro lado, já tinha também argumentado em Santos (1996) que a escolha dos textos e da informação a eles associada tinha de ter um objectivo (senão, estaríamos apenas em presença de uma colecção). Em 2006, tive a oportunidade de, no contexto didáctico da Primeira Escola de Verão da Linguateca (Santos, 2006), produzir uma definição mais precisa e mais abrangente do que me parece serem os factores constitutivos deste instrumento, o corpo (linguístico), que passo a citar aqui:

> Um corpo é uma colecção classificada de objectos linguísticos para uso em Processamento de Linguagem Natural/Linguística Computacional/Linguística

em que *uso* pode ser estudo, medição, teste, ou avaliação, enquanto variados *objectos linguísticos* são textos, frases, palavras, entrevistas, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, filmes, imagens com legendas, traduções, correcções (de textos de alunos de língua ou de tradução), telefonemas, simulações do

---

[2] Por si só, esta designação já implica uma grande falta de rigor, visto que o que os seus adeptos defendem é a "Web como fonte de corpos" ou, ainda de forma mais longínqua, a "Web como informante".

tipo Wizard of Oz, programas... Para exemplos destes tipos diferentes de corpos, veja-se o material da Escola. Por seu lado, a palavra *classificada* pode referir-se a muitas questões diferentes:

- A nível dos parâmetros da recolha: que categorias considerar;
- A nível da escolha: todos, alguns, amostra,... (Santos (2000), Mair (1992);
- A nível dos fenómenos: tipo de erro, tipo de tradução, tipo de texto, ...
- A nível dos constituintes: análise sintáctica, semântica, fonológica, discursiva, etc.
- Avaliação (quando existem julgamentos associados)

Contudo, o mais importante num corpo é saber o que fazer com ele, como usá-lo, e para que tarefas ele é útil. (Outra questão, relacionada, será a necessidade de criar um novo corpo se não houver nenhum apropriado para as nossas demandas.) É sobre esse assunto que pretendo dedicar maioritariamente este texto, e – porque não? – tentando pôr as pessoas a aproveitar os corpos que já existem em vez de compilar cada uma o seu.

O meu ponto de partida é o de que um corpo não é o objecto de estudo do que em inglês se chama *corpus linguistics*, mas sim a ferramenta, o utensílio com que se faz linguística, por isso a minha denominação "linguística com corpos". Não posso discordar mais da afirmação de Kilgarriff (2001) quando afirma que um corpo é o objecto de estudo da CL/LC, e que cito integralmente em seguida:

> There is a void at the heart of corpus linguistics. The name puts 'corpus' at the centre of the discipline. *[nota:* Alternative names for the field (or a closely related one) are "em-

> pirical linguistics" and "data-intensive linguistics". By using an adjective rather than a noun, these seem not to assert that the corpus is an object of study. Perhaps it is equivocation about what we can say about corpora that has led to the coining of the alternatives./ In any science, one expects to find a useful account of how its central constructs are taxonomised and measured, and how the subspecies compare. But to date, corpus linguistics has measured *corpora* only in the most rudimentary ways, ways which provide no leverage on the different kinds of corpora there are.

Na minha opinião, isto é o mesmo que dizer que os cadáveres em medicina, ou os ratos de laboratório, em farmácia, são o objecto de estudo destas disciplinas. Não, eles são formas de estudar o corpo humano e o metabolismo, mas nunca o objecto de estudo.

Por isso, e feito este esclarecimento, vejamos um corpo (que é um objecto finito e concreto) como um utensílio para estudar a língua (ou literatura ou cultura).[3]

Um preceito que reputo de essencial é, assim, que a primeira coisa que convém tornar clara, é o que se pretende saber sobre uma língua, e só depois, muito depois, como é que um corpo nos pode ajudar. Como por várias vezes já mencionado – veja-se Sankoff (1978) ou Svartvik (2007) – não é óbvio que seja preciso recolher frases ou textos de outros autores para estudar a própria língua... como é preciso recolher, necessariamente, exemplares de folhas e flores para estudar botânica, por exemplo. Por outro lado, existe extensa literatura a fundamentar as necessidades e vantagens de usar material externo ao próprio linguista, não só para domínios de estu-

[3] Veja-se também a posição tão bem exposta por Chafe (1992) dos perigos de identificar a linguística com um instrumento (entre vários).

do que se encontram por assim dizer também "fora" do falante individual, como a dialectologia, o estudo das línguas estrangeiras, a aquisição da língua por crianças ou a diacronia, mas também para a sua própria língua materna.

Com efeito, exteriorizar o material de estudo permite outras visões, outras opiniões, e a comparação com outros falantes, além de nos ajudar a identificar problemas e consciencializar-nos de aspectos de que não estaríamos conscientes. Para além disso, uma questão muito importante que os corpos trazem à linguística como actividade científica (Santos & Oksefjell, 1999) é a impossibilidade de viciar a análise no sentido de produzir exactamente as frases que dariam jeito para uma dada teoria (inconscientemente, claro), e – talvez o mais importante de tudo – a possibilidade de quantificar. Com efeito, uma das mensagens que salientarei, mais à frente, é a importância da distribuição e não apenas da concordância.

Há, contudo, duas observações que se impõem em relação à quantificação: a primeira, é que a linguística quantitativa não é necessariamente baseada em corpos (basta apreciar o índice das revistas respectivas); por outro lado, a linguística com corpos é maioritariamente qualitativa ou ilustrativa. Embora esse último carácter provenha de haver muitas publicações nesta área com o único ou principal fim de (i) descreverem recursos e aliciar leitores, alunos e professores para o seu uso (muitas vezes indiscriminado), ou (ii) relatar experiências concretas de usos de corpos, por exemplo na sala de aula, existem também demasiadas obras, na minha opinião, que se limitam a apresentar resultados ou valores sem qualquer preocupação de explicar porque é que a recolha desses valores tem importância ou interesse.[4]

---

[4] Penso que todos os leitores já se depararam com esta situação, sendo inútil invocar casos específicos.

Tipos de estudos com corpos

Parece-me muito importante começar por fazer a distinção entre dois tipos de estudos empíricos: *exploratórios* e *experimentais*. Em ambos, os corpos podem ter um papel fundamental. Esta distinção é bem conhecida nas ciências empíricas (ciências da natureza e ciências sociais, Cohen (1995)), mas aparentemente ainda não é do domínio geral no campo da linguística...

Um estudo *exploratório*, como o seu nome indica, procura coisas interessantes para mais tarde estudar. Colige amostras, conta ocorrências, surpreende-se com casos que se deparam ao investigador. Procura correlações, experimenta classificações, identifica conjuntos. Por outras palavras, abre sendas, identifica lugares de interesse (para lá voltar ou para outros lá irem). Tecnicamente, constrói uma teoria ou um mapa da área.[5]

Um estudo *experimental*, por outro lado, já tem uma hipótese ou conjunto de hipóteses que pretende verificar. Uma hipótese, para ser digna desse nome, é algo que extravasa o *corpus* mas se refere à língua (ou à cultura), e que possa ser confirmado empiricamente de forma indirecta. Por exemplo: *que há mais verbos X do que verbos Y* não é uma hipótese: é um dado, que se pode verificar (ou não) num dado corpo. *Que a língua privilegia a expressão abstracta*, ou *que pode ser descrita por um certo modelo W*, já são hipóteses cujas consequências concretas se podem aferir. Ao contrário do que seria talvez esperável (e certamente desejável), quanto mais precisa a hipótese (estatística), mais dados são precisos para a testar.[6]

---

[5] Veja-se Gardenfors (2000) para a metáfora da geografia na conceptualização, que ele usa exemplarmente.

[6] Por outras palavras, o tamanho de corpos suficientes para produzir bons estudos exploratórios é muito menor do que o requerido por estudos que exijam significância estatística.

Claro está que, na prática, a maior parte dos estudos têm uma componente exploratória e outra experimental. Além disso, um estudo experimental é geralmente produzido com base nas explorações de outros.[7] Apresento dois exemplos concretos dos dois tipos de estudos, remetendo o leitor para eles para mais pormenores:

Em Inácio et al. (2008) resolvemos explorar o domínio da cor, procurando medir todas as dependências ou correlações entre factores que um *corpus* paralelo nos podia oferecer. Em alguns casos, não conseguimos obter nenhuma regularidade, noutros, o estudo forneceu-nos pistas para formular generalizações, se bem que incipientes.

Em Santos (2008), tentei confirmar a hipótese de que a língua portuguesa não gramaticaliza (nem exprime, na maior parte das vezes) o resultado, ao contrário da língua inglesa que tem essa categoria como gramatical (no *perfect*). Aqui parti de uma hipótese precisa e tentei gizar um conjunto de previsões que pudessem ser verificadas num corpo bilingue.

Corpos para outros objectivos

Há também que indicar que muita (senão a maioria da) actividade feita com corpos não é, de facto, estritamente linguística (no sentido de estudar a língua), mas sim "aplicada", no sentido de construir dicionários (ou tesauros, ou gramáticas); ou no sentido de testar aplicações de processamento de linguagem natural (ou recolha de informação, RI).

---

[7] Embora haja também uma outra abordagem dos estudos experimentais com corpos: simplesmente pegando numa teoria em que se acredita (ou não) e mostrando como permite (ou não) prever certas observações.

De facto, a palavra corpo (ou *corpus*) ainda continua a ser usada nos nossos dias com dois sentidos diferentes: o linguístico, definido acima, e o informático, no sentido de material de teste e de treino, veja-se Sparck-Jones & Galliers (1996) sobre as chamadas colecções de teste em RI. Estritamente, também se devia ainda discriminar o uso do corpo como matéria-prima (para construir, por exemplo, diconários ou materiais de ensino).

Sistematizando, podem identificar-se quatro tipos de usos de corpos:

1. Em primeiro lugar, usa-se um corpo para ter uma ideia do problema/conhecer, dando origem às metáforas do "corpo como consultor", "corpo como familiarizador", "corpo como treinador", ou "corpo como mar de língua".
2. Em segundo lugar, usa-se um corpo para medir um dado fenómeno.
3. Em terceiro e mais comummente, para avaliar algo: uma hipótese, um sistema, um método, uma teoria...
4. Finalmente, o uso talvez mais frequente é para criar outras coisas, e entre estas destaco: a) dicionários ou outras estruturas de conhecimento, como terminologias, almanaques e ontologias, b) materiais de teste de ensino de línguas, c) sistemas de resposta automática a perguntas (RAP), d) sistemas de ensino, e) jogos, f) sistemas de detecção de plágio, de correio não endereçado (*spam*), ou outros[8]

[8] Não me posso alongar aqui na teoria da avaliação, mas é fundamental separar conceptualmente os corpos usados na criação, desenvolvimento, ou mesmo incluídos dentro dos sistemas mencionados, dos corpos construídos como recursos dourados para a própria avaliação desses mesmos sistemas, ou seja, materiais de avaliação cobertos pela alínea anterior.

Naturalmente, nem todos os corpos são apropriados para todos os usos. Por outro lado, embora muitos corpos não sejam criados com apenas um destes objectivos (pelo contrário, pretendem ser de uso geral, ou suficientemente geral), há um compromisso incontornável entre o desenho e o tipo de usos de um corpo. Assim como todos os instrumentos que são criados para muitos usos não são nunca tão afiados como aqueles que foram desenhados para um uso particular, temos na criação de corpos este conflito universal que opõe generalidade a optimização.

Metodologia da linguística com *corpora*

Voltemos então à questão de como escolher um corpo, ou melhor, como escolher as perguntas/estudos/anotação de um corpo já existente para responder ou investigar uma dada necessidade de informação.

Ao dar primazia ao uso ou emprego dos corpos, não estou de forma alguma a desprezar o trabalho de bastidores, e sobretudo de documentação, que é preciso fazer para que os utilizadores possam apreciar o que estão a consumir/usar... como aliás o demonstra o trabalho que temos tido na Linguateca na produção de documentação e de corpos com qualidade.[9] E, também, ao constatar repetidamente que os utilizadores usam mal ou desajustadamente os corpos que pomos à disposição deles.

---

[9] Por exemplo, a questão da revisão posterior dos textos electrónicos que foram incorporados no COMPARA, e que nos levaram ao confronto de múltiplas edições da mesma obra.

Sou da opinião, aliás, de que muitos projectos de corpos falham precisamente neste aspecto, acabando por "vender gato por lebre", quando por exemplo dão ao utilizador uma série de instrumentos, aparentemente muito científicos, tais como medidas de associação, ou análise gramatical, sem tornar claro os casos em que tais medidas são aplicáveis ou sem referir minimamente quais os fundamentos gramaticais (a todos os níveis) empregues. Porque mesmo a acção mais simples imaginável, a de contar palavras ou identificar a pontuação, pressupõe uma teoria linguística (Grefenstette & Tapanainen, 1994, Nunberg, 1990), ou, na sua ausência, uma descrição detalhada de todos os casos cobertos pelo corpo (Sampson, 2003).

Vejamos um exemplo: enquanto o *Corpus* do Português[10] de Mark Davies apresenta uma interface realmente bem desenhada e imaginada, oferecendo muitas possibilidades diferentes ao utilizador, o conteúdo que serve, pelo contrário, foi alvo de um processamento muito pouco cuidado, que levará, na minha opinião, a estudos com pouco fundamento linguístico. De facto, não existe em todo o sítio uma única informação sobre a forma como a categoria gramatical foi atribuída. Contraponho aqui o trabalho que investimos na Linguateca na revisão humana da anotação gramatical, na Floresta Sintá(c)tica e no COMPARA, com extensiva documentação das opções e do que as anotações realmente significam.[11]

---

[10] http://www.corpusdoportugues.org/

[11] http://www.linguateca.pt/Floresta e http://www.linguateca.pt/COMPARA

Complementaridade, ao invés de igualdade, da anotação e do léxico

Uma outra confusão que me parece infelizmente bem expandida é a de igualar (ou não distinguir) entre análise de contexto e conhecimento lexical. Por exemplo, Leech (1993) afirma que as categorias morfossintácticas que devem estar num corpo anotado são as dos léxicos das línguas respectivas. Ora, exactamente o que um corpo pode trazer é informação sobre algo que ainda não está no léxico, nem pode estar, sem despir o conceito de léxico de todo o sentido: por exemplo, o facto de um adjectivo ser usado como nome[12], o facto de uma dada construção acontecer mais frequentemente com um dado tempo verbal, ou a constatação de que um verbo co-ocorre maioritariamente com participantes humanos. O corpo é para investigar a língua em contexto, enquanto que o dicionário cobre e fixa aquilo que é inerente aos itens lexicais independentemente do contexto.[13]

Distribuição vs. concordância

Mais importante do que o número de vezes que um dado fenómeno ocorre, ou a observação do mesmo, é a informação – que só um corpo pode dar – do peso relativo de uma dada questão em relação ao universo representado. Quem ainda não interiorizou esta distinção é um utilizador muito ingénuo de um corpo...

---

[12] Por outras palavras, "ser usado em posição nominal" (categoria do corpo) não é o mesmo do que ser classificado como substantivo (categoria do léxico).

[13] Evidentemente, estas linhas não podem passar de uma simplificação grosseira sobre a delicada problemática do balanço entre o léxico e a gramática, que está no cerne de muito debate em linguística. Mas o meu objectivo aqui era tão só realçar a complementaridade, ao invés da igualdade.

Ou seja, a frequência absoluta de um dado fenómeno é completamente ininterpretável sem relação com o número máximo de casos possíveis quando essa contagem foi efectuada. Já a frequência relativa (que é o quociente do número de ocorrências pelo número total) junto com a distribuição por diferentes categorias permitem uma primeira noção sobre a importância e a correlação com estas últimas.

Propriedades estatísticas

Uma propriedade interessante da língua é o facto de haver sempre muitos casos diferentes com pouca frequência, enquanto os casos de alta frequência corresponderem sempre a poucos casos (matematicamente falando, a ordem e a frequência são inversamente proporcionais). Esta propriedade é denominada por lei de Zipf, e embora Zipf (1949) tenha proposto esta "lei" para praticamente toda a actividade humana, em linguística esta regularidade costuma ser ilustrada relativamente às palavras que ocorrem num texto (ou num corpo).

Mas deixem-me explicar exactamente o que significa esta lei, com ajuda de figuras e da sua definição formal. Graficamente, se ordenarmos de forma decrescente um conjunto de fenómenos pela sua frequência, obtemos uma função como a representada na figura 1[14] (ou na figura 2, em escala logarítimica em ambos os eixos).

[14] É importante salientar que este gráfico representa uma função matemática contínua, com valores para todos os casos, enquanto que as observações sobre os valores da frequência de um dada palavra ou construção são sempre discretos. A figura 2 é mais correcta na identificação dos pontos reais.

$$O(i) = \frac{n}{i^{\Theta} H_{\Theta}(V)}$$

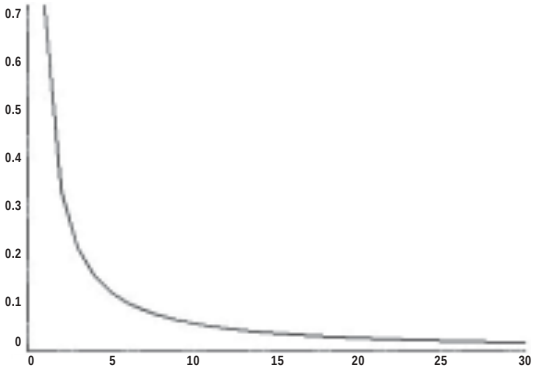

Figura 1. Representação gráfica do número de ocorrências O em função da ordem i obtida em termos de frequência. n e $H_{\Theta}(V)$ são constantes relacionadas com o tamanho das observações (ou dados) e do vocabulário V. Figura retirada de: http://planetmath.org/encyclopedia/ZipfsLaw.html

De uma forma rigorosa, a lei de Zipf pode enunciar-se assim: a frequência de ocorrência de um dado acontecimento (palavra, construção, etc.) **O** é uma função – da forma $1/i^{\Theta}$ (em que o expoente **Θ** é próximo da unidade) – da ordem ("rank") **i**, quando essa ordem é estabelecida em termos da frequência de ocorrência. Mais simplificada e aproximadamente, o produto da frequência pela ordem é uma constante (esta formulação seria exacta se o expoente fosse 1): **i x O(i) = C**

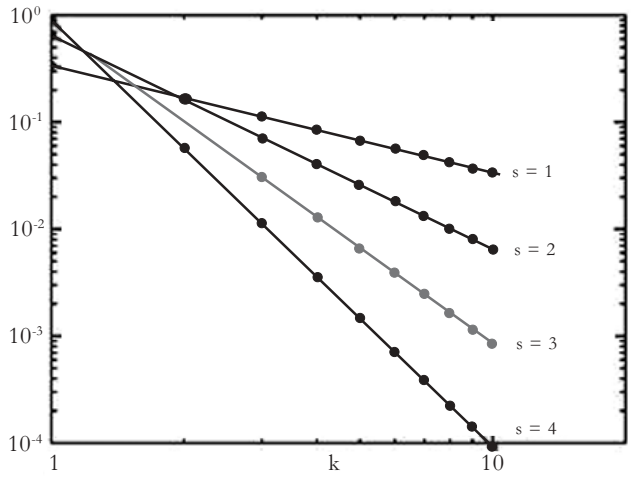

Figura 2. Representação gráfica da lei de Zipf em escala logarítmica, com n (número de observações) = 10, e k representando a ordem. s representa diferentes constantes. Figura retirada de http://en.wikipedia.org/wiki/Zipf%27s_law

Ou seja, se ordenarmos um conjunto de observações (por exemplo palavras) pela sua frequência, atribuindo à mais frequente a ordem 1, à segunda mais frequente a ordem 2, e por aí adiante... o produto da ordem pela frequência mantém-se quase constante. De outra forma ainda: a frequência da segunda observação (palavra) é metade da primeira, a da terceira é um terço da primeira...

A minha mensagem neste artigo é que esta lei (ou regularidade) é válida e observável também se o nosso alvo de análise for a distribuição quantitativa de construções sintácticas, lemas, traduções (de facto, qualquer variável linguística que se possa contar), o que significa que: a) teremos de a ter em conta ao tirar conclusões sobre os nossos dados quantitativos, assim como b) ela nos dá logo de princípio algo que podemos prever.

Para tornar mais claro o que significa a lei de Zipf aplicada a outros fenómenos que não apenas a frequência das palavras, veja-se a tradução de tempo verbal entre o português e o inglês. No pequeno corpo estudado na minha tese (Santos, 2006, p. 328) tínhamos por exemplo os seguintes valores para a tradução dos 2305 casos de "simple past", repetidos na Tabela 1 (os mais frequentes):

Perfeito	1135
Imperfeito	913
Infinitivo	57
Mais que perfeito	39
Gerúndio	35
Imperfeito conjuntivo	26
Particípio passado	18
Condicional	14
Presente	14
ir + gerúndio	12
outros cinco	29

Tabela 1: tempo verbal na tradução do simple past inglês

Este exemplo é sintomático da aproximação zipfiana: o corpo é demasiado pequeno para a lei de Zipf dar números fiáveis,[15]

[15] De notar que, sendo uma lei empírica, os números obtidos serão sempre uma aproximação. Quanto mais dados mais próxima fica da forma teórica. De notar também que a lei de Zipf é uma família de leis (conforme os coeficientes).

mas permite prever que, se juntarmos mais casos de tradução, iremos encontrar ainda mais casos diferentes com menor frequência, aumentando os casos já encontrados.[16]

Outra regularidade estatística que tem sido recentemente discutida e que é aparentada com a lei de Zipf é a chamada cauda comprida (em inglês, *long tail*), muito em voga na economia e pretendendo explicar o peso de coisas "leves", tal como livros lidos por poucas pessoas mas representando em conjunto metade das vendas, veja-se Kilkki (2007). Uma analogia útil para linguística com corpos é o facto de muitas questões raras se tornarem a quase maioria da língua (e não se poderem portanto desprezar): praticamente para qualquer frase que se olhe, ou analise, é possível encontrar algo pouco frequente ou raro... mas, se retirássemos esses casos, ficaríamos sem texto!

A esse propósito, uma questão muito pertinente, discutida por Gale & Sampson (1995), é a de como estimar números de ocorrências de casos (tipos) ainda não encontrados no material. Esta é obviamente uma consideração essencial: se, pela lei de Zipf, quanto maior um corpo, maior o número de casos raros que surgem, não se pode admitir que todas as construções – ou casos – estejam lá, por muito grande que seja o corpo. Existem assim já alguns métodos desenvolvidos para lidar com esta questão na própria linguística.

Outra perspectiva sobre a qual a estatística linguística se tem debruçado é a da distribuição de fenómenos ao longo de um texto

---

[16] Halliday (2005) tem a seguinte visão da distribuição estatística dos fenómenos gramaticais: há apenas dois tipos de distribuição 0,5:0,5 e 0,1:0,9, esta última modelando as categorias não marcadas. A interacção deste modelo com a lei de Zipf poderá explicar melhor os números da tabela.

(note-se que a lei de Zipf se refere ao conjunto não ordenado dos fenómenos). Nessa óptica, outras regularidades têm sido investigadas, nomeadamente a diferença de comportamento entre as palavras gramaticais (preposições, verbos auxiliares, artigos, conjunções, etc.) e as palavras plenas (substantivos, verbos, adjectivos, nomes próprios, etc.). Com efeito, enquanto as primeiras tendem a aparecer uniformemente ao longo de um texto, as segundas vêm "aos soluços", com concentrações locais e não globais: ou seja, se um nome próprio aparece num texto, prevemos que apareça mais vezes nesse mesmo texto e talvez até próximo... e não apenas vinte páginas depois (KATZ, 1996). Esse tipo de comportamento permite métodos de detecção do conteúdo, como o sugerido por Scott (2006) na detecção de palavras-chave de um texto.

As linhas acima são, evidentemente, uma iniciação muito ligeira às questões estatísticas, que são tão ou mais relevantes para a linguística com corpos do que para a linguística computacional (em que a estatística é muitas vezes simplesmente uma ferramenta). Gostava a este respeito de chamar a atenção para o comentário irónico de Gale & Sampson (1995), após apresentarem pormenorizadamente uma dada técnica: *However, applications of the technique are likely to be more judicious when based on an awareness of its rationale*. Uma tradução mais directa seria: "se não percebe o que está a fazer, não se aventure!" Não me parece que uma postura de "eu não percebo disso, a minha formação é outra..." possa ser invocada na pesquisa: Se queremos compreender, não podemos fixar-nos ou fiar-nos em áreas estanques, mas sim estar abertos ao diálogo e à compreensão de métodos e conceitos aparentemente de "outras" áreas. "Área" é aliás algo dinâmico que está sempre em evolução...

Alguns temas interessantes

Por fim, gostava de mencionar alguns assuntos que me parecem muito interessantes para estudar em linguística com corpos, e que têm a ver com a forma como a língua (no nosso caso, o português) é usado.

O primeiro é o humor: o que é que tem graça, quais as formas de ser criativo, irónico, engraçado em português (ou agressivo, mal-disposto, sarcástico, cáustico...) Quanto do humor está ligado à cultura, e quanto está ligado à língua?[17]

Estou convencida que foi um erro fulcral a separação entre literatura e linguística, no sentido de que devia ser (também) a estudar a literatura que compreendíamos a língua. A divisão positivista entre informação objectiva e tudo o resto não tem qualquer razão de ser, mas ainda reina, aparentemente, na maioria dos departamentos de Letras, separando os estudos literários da linguística. Veja-se Ellis (1993) para uma visão original e radical da língua em que os elementos básicos são emotivos/afectivos e não "objectivos". Comungando da mesma crença, acho que em vez de afastar a linguagem literária das nossas lupas, devíamos aceitá-la e estudá-la como modelo arquetípico que é para a(s) nossa(s) cultura(s).[18]

---

[17] Como Stella Tagnin comentou, língua e cultura não são separáveis, e daí a importância, também, do estudo das diferentes variantes: pois algo pode ser humor intencional por parte do autor, ou ser ridículo por causa de diferenças linguísticas. Por outro lado, quanto mais longe (em tempo ou em grau de estranheza) se encontrarem duas culturas mais difícil é apreciar o humor (ou mesmo dar por ele), independentemente da língua em que está formulado.

[18] Aliás, estou convencida de que a intertextualidade cruza géneros, ou seja, a criatividade é apanágio, e utensílio, de todo o bom redactor enquanto criador de textos.

O segundo é a relação da língua com as imagens: afinal de contas, e contradizendo a sabedoria popular, uma palavra vale mais do que mil imagens! Para nos apercebermos disso, basta tentar arranjar imagens para conceitos como "desonestidade", "graça", "mentira"... Na realidade, existem relações muito interessantes entre imagens e texto, cada um emprestando sentido ao outro. Num mundo cada vez mais multimédia, o estudo da relação entre estes dois modos impõe-se tanto numa perspectiva de criação de conteúdos como de recuperação dos mesmos.

Finalmente, estudos de tradução como uma terceira língua, ou seja, assumindo a subalternidade do "português traduzido" em relação ao puro/original, parecem-me profundamente errados, além de minimizadores de um grupo dos mais criativos: os tradutores (Santos, 2007). Na minha óptica, seria mais interessante prosseguir uma comparação de diferentes estratégias e interesses diferentes provocados por línguas diferentes, desde questões de pormenor (micro-nível) como classes aspectuais (Santos, 2004), a questões mais gerais (macro-nível) como a descrição de personagens masculinas ou femininas.

Agradecimentos

Agradeço a Stella Tagnin o amável convite para participar neste volume; agradeço a Belinda Maia a parte de leão na organização da Primeira Escola de Verão da Linguateca, em 2006, na qual parte deste material foi pela primeira vez organizado e apresentado; agradeço a Stig Johansson o apoio e ajuda que deu na minha corporização, desde lugar para trabalhar na Universidade de Oslo até bibliografia relevante e correcção dos meus textos em inglês macarrónico. Finalmente, se não fosse o desafio que Lauri Carlson me

lançou, de testar as minhas teorias em corpos paralelos durante o doutoramento, eu não estaria certamente a escrever este artigo.

Este artigo foi produzido no âmbito da Linguateca, contrato número 339/1.3/C/NAC, financiado pelo governo português e pela União Europeia.

Referências

CHAFE, Wallace. The importance of *corpus* linguistics to understanding the nature of language. In Jan Svartvik (ed.), *Directions in* Corpus *Linguistics: Proceedings of Nobel Symposium 82* (Stockholm, 4-8 August 1991), pp. 79-97.

COHEN, Paul R. *Empirical Methods for Artificial Intelligence*. The MIT Press, 1995.

ELLIS, John M. *Language, Thought and Logic*. Evanston IL: Northwestern University Press, 1993.

FACCHINETTI, Roberta (ed.) Corpus *Linguistics 25 Years on*. Rodopi, 2007.

GALE, William A. & Geoffrey SAMPSON. Good–Turing Frequency Estimation Without Tears, *Journal of Quantitative Linguistics 2*, 1995, pp. 217–37.

GARDENFORS, P. *Conceptual Spaces: The Geometry of Thought*. Cambridge: The MIT Press, 2000.

GREFENSTETTE, Gregory & Pasi TAPANAINEN. What is a word, What is a sentence? Problems of Tokenization, *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX'94),* pp. 79-87.

HALLIDAY, M.A.K. *Computational and Quantitative Studies*, vol 7. In the Collected Works of MAK Halliday, edited by Jonathan J. Webster. London & New York: Continuum, 2005.

INÁCIO, Susana; Diana SANTOS & Rosário SILVA. COMPARAndo cores em português e inglês. In Sónia Frota & Ana Lúcia Santos (eds.), *Textos seleccionados apresentados ao XXIII Encontro da Associação Portuguesa de Linguística* (Évora, 1-3 de Outubro de 2007), APL, 2008.

KATZ, Slava M. Distribution of content words and phrases in text and language modelling, *Natural Language Engineering 2* (1996), pp.15-59.

KILGARRIFF, Adam. "Comparing *corpora*", *International Journal of Corpus Linguistics 6* (1), 2001, pp. 1-37.

KILGARRIFF, Adam & Gregory GREFENSTETTE. Introduction to the Special Issue on Web as *Corpus*. *Computational Linguistics 29* (3), 2003, pp. 333-348.

KILKKI, Kalevi. A practical model for analyzing long tails, *First Monday 12*, 5, May 2007, http://www.firstmonday.org/issues/issue12_5/kilkki/

LEECH, Geoffrey. *Corpus* Annotation Schemes. *Literary and Linguistic Computing 8* (1993), pp. 275-81.

MAIR, Christian. Comments to Wallace Chafe's paper. In Jan Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82* (Stockholm, 4-8 August 1991), pp.98-103.

MATOS, Francisco Gomes de. O Cientista de Língua Portuguesa e seus Direitos Linguísticos, *Revista Internacional de Língua Portuguesa 7,* 1992, pp. 79-81

NUNBERG, Geoffrey. *The linguistics of punctuation,* CSLI Lecture Notes, Number 18, 1990.

OKSEFJELL, Signe & Diana SANTOS. Breve panorâmica dos recursos de português mencionados na Web. In Vera Lúcia Strube de Lima (ed.), *Anais do Terceiro Encontro de Processamento da Língua Portuguesa (Escrita e falada), PROPOR'98* (Porto Alegre, 3-4 novembro 1998), pp. 38-47.

SAMPSON, Geoffrey. Thoughts on two decades of drawing trees. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora,* Kluwer Academic Publishers, 2003, pp. 23-41.

SANKOFF, David. Probability and linguistic variation, *Synthese 37* (1978), pp. 217-238.

SANTOS, Diana Maria de Sousa Marques Pinto dos. Tense and aspect in English and Portuguese: a contrastive semantical study, Tese de doutoramento, Instituto Superior Técnico, Universidade Técnica de Lisboa, Junho de 1996.

SANTOS, Diana. Providing access to language resources through the World Wide Web: the Oslo *Corpus* of Bosnian Texts. In Antonio Rubio, Natividad Gallardo, Rosa Castro and Antonio Tejada (eds.), *Proceedings of The First International Conference on Language Resources and Evaluation (*Granada, 28-30 May 1998), Vol. 1, pp.475-481.

SANTOS, Diana. Processamento computacional da língua portuguesa: Documento de trabalho. 1999, http://www.linguateca.pt/branco/index.html

SANTOS, Diana. The translation network: A model for the fine-grained description of translations. In Jean Véronis (ed.), *Parallel Text Processing,* Dordrecht: Kluwer Academic Publishers, 2000, pp.169-186.

SANTOS, Diana. Um centro de recursos para o processamento computacional do português, *DataGramaZero – Revista de Ciência da Informação* v.3 n.1 fev/02, http://www.dgz.org.br/fev02/Art_02.htm.

SANTOS, Diana. *Translation-based corpus studies: Contrasting English and Portuguese tense and aspect systems.* Amsterdam/New York, NY: Rodopi, 2004.

SANTOS, Diana. Desenho, construção e utilização de *corpora*. *Material de ensino na Primeira Escola de Verão da Linguateca* (Universidade do Porto, Portugal, 10 de Julho de 2006), http://www.linguateca.pt/escolaverao2006/*Corpora*/*Corpora*EscolaVerao.pdf.

SANTOS, Diana. A tradução na sociedade do conhecimento OU Tradução: uma tecnologia humana de ponta OU Ciência E Tradução. In *Actas do IX Seminário de Tradução Científica e Técnica em Língua Portuguesa* (Lisboa, 13 de Novembro de 2006), Lisboa: União Latina, 2007, CD-ROM.

SANTOS, Diana. Perfect mismatches: Result in English and Portuguese. In Margaret Rogers & Gunilla Anderman (eds.), *Incorporating Corpora: The Linguist and the Translator***.** Clevedon: Multilingual matters, 2008, pp. 217-242.

SANTOS, Diana & Signe Oksefjell. Using a Parallel *Corpus* to Validate Independent Claims, *Languages in contrast 2* (1), 1999, pp.117-132.

SCOTT, Mike. Key words of individual texts: Aboutness and style, Chapter 4 of Scott, Mike & Christopher Tribble. *Textual Patterns: Keywords and corpus analysis in language education*. Amsterdam/Philadelphia, Benjamins, 2006, pp. 55-72.

SPARCK-JONES, Karen & Julia R. GALLIERS. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer, 1996.

SVARTVIK, Jan. *Corpus* linguistics 25+ years on. In Facchinetti, Roberta (ed.), *Corpus Linguistics 25 Years on*. Rodopi, 2007, pp. 11-25.

ZIPF, George Kingsley. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, Mass.: Addison-Wesley Press , 1949.